STAFF SUMMARY SHEET

	TO	ACTION	SIGNATURE (Surname), GRADE AND DATE		TO	ACTION	SIGNATURE (Surname), GRADE AND DATE
1	DFE	sig	Enger AD-25 11 Sep 14 (Department Head or Designee)	6			
2	DFER	approve	Solt, AD22, 25 Sept 14	7			
3	DFEI	action	(Author /Originator)	8			
4				9			
5				10			

SURNAME OF ACTION OFFICER AND GRADE	SYMBOL	PHONE	TYPIST'S INITIALS	SUSPENSE DATE
Thomas, Lt Col	DFEI	333-0673	JYT	20141031

SUBJECT
Clearance for Material for Public Release USAFA-DF-PA- 440

DATE
20140822

SUMMARY

1. PURPOSE. To provide security and policy review on the document at Tab 1 prior to release to the public.

2. BACKGROUND.
Authors: Lt Col Joseph Y. Thomas & Dr David P. Diros (Oklahoma State University)

Title: Theoretical Validation of IDT in Real-World, High-Stakes Deceptive Speech
Circle one: Journal Article

Description: The study of deception and the theories which have been developed have relied heavily on laboratory experiments, in controlled environments, utilizing American college students, participating in mock scenarios. The goal of this study was to validate previous deception research in a real-world high-stakes environment. This study utilized previously confirmed speech cues and constructs to deception in an attempt to validate a leading deception theory, Interpersonal Deception Theory (IDT). The results did validate IDT with mixed results on individual measures and their constructs.

Release Information: (#1) Conference Symposium: Hawaii International Confernece On Systems Sciences, Symposium: Rapid Screening Technologies, Deception Detection and Credibility Assessment. Leading to (#2) journal publication in: the Journal of Information Science.

Recommended Distribution Statement:
(Distribution A, Approved for public release, distribution unlimited.)

3. DISCUSSION.

4. VIEWS OF OTHERS.

5. RECOMMENDATION. Department Head or designee reviews as subject matter expert. DFER reviews for policy and security. Coordination indicates the document is suitable for public release. Suitability is based on the document being unclassified, not jeopardizing DoD interests, and accurately portraying official policy [Reference DoDD 5230.09]. Release is the decision of the originator (author). Compliance with AFI 35-102 is mandatory.

Joseph Y. Thomas, Lt Col, USAF
DFM Instructor, IITA Director IT Research

1 Tabs
1. copy of journal article

Theoretical Validation of IDT in Real-World, High-Stakes Deceptive Speech

Lt Col Joseph Y. Thomas
Institute for Information Technology Application
United States Air Force Academy
Colorado springs, CO
Joseph.thomas@usafa.edu

Dr David P. Biros
Dept of Management Science and Information systems
Oklahoma State University
Stillwater, OK
David.biros@okstate.edu

***Abstract*—The study of deception and the theories which have been developed have relied heavily on laboratory experiments, in controlled environments, utilizing American college students, participating in mock scenarios. The goal of this study was to validate previous deception research in a real-world high-stakes environment. This study utilized previously confirmed speech cues and constructs to deception in an attempt to validate a leading deception theory, Interpersonal Deception Theory (IDT). The results did validate IDT with mixed results on individual measures and their constructs.**

Keywords—deception; real-world; high-stakes; speech

I. INTRODUCTION

Deception is a ubiquitous form of communication [1]. In fact deception is a major characteristic of the most common communication channels; 14% of people self-reported deceiving in emails, 37% in phone calls, and 27% in face-to-face interactions [2]. The formal study of deception detection and its cues has been covered in numerous cross discipline studies and the consensus across the board is that humans are poor detectors of deceit [3], [4], [5]. In perhaps the most comprehensive meta-analysis of deception detection cues and their accuracy, Bond and DePaulo [3], looked at 206 studies with 24,483 judgments and found a mean accuracy of 53.4%. To be more colloquial, humans might as well flip a coin when it comes to detecting deception. However, humans are not just inaccurate detectors of deceit but poor judges of what cues are indicators of deception and are often affected by multiple biases [6]. Human bias toward unreliable deceptive cues hampers our ability to perceive deception and can further decrease accuracy below chance [6]. Therefore humans have long searched for behaviors and tools to aid them in detecting deceit.

Current methods to detect deception all have drawbacks and can be split into two categories, invasive and non-invasive. Of the invasive technologies currently available to help identify and measure deceit, the polygraph is the most well-known. The polygraph is a device that takes various cardiac, skin conductivity, and respiratory measures to detect deception. It is based on the idea that these physiological measures are directly linked to the conditions that are brought on by deception attempts [6]. In a summary of laboratory tests, Vrij reports that the polygraph is about 82% accurate at identifying deceivers [6]. However, polygraph exams have several strong limiters namely a willing subject, an invasive exam, and the need for a trained examiner. The polygraph exam itself can evoke fear and apprehension in its subjects making it a controversial investigative tool.

The newest invasive method to detect deception utilizes functional magnetic resonance imaging (fMRI) to map blood flow in the brain during structured questioning. One fMRI study reached deception detection accuracy of 100% when subjects do not employ countermeasures [7]. fMRI measures the hemodynamic response, or changes in blood flows, that are related to brain activity. Researchers have noticed differences between the brain activity of truth-tellers and deceivers [8], [9], [10]. Though initial findings are promising, fMRI shares the same restrictions as the polygraph (a willing subject, an invasive exam, and the need for a trained examiner). Additional limiters to their general use are their sheer size their cost to operate, the fact that subjects cannot move at all, and they cannot be used on people with claustrophobia or metallic implants. These leading deception detection tools are prohibitive [11] and emphasize the need for less obtrusive means of measuring deceptive behavior that do not require human intervention.

In addition to prohibitive tools, the current methodology used in most deception detection research is lacking in areas that separates it from real-world settings [12]. A vast majority of current deception detection research utilize American university students instructed to lie in mock scenarios [12] [13]. But research in high-stakes environments, such as interviews during a criminal investigation, is deficient [14], [12], [15], [16], [17], [18]. This has driven a strong need for more field studies in deception detection research [19].

A principal deception detection meta-analysis of 120 studies showed 101 used student subjects [12]. Only four of these studies (3%) involved situations where the subjects were not given instructions to lie but chose to do so on their own. There is evidence that behavior differs between those who choose to lie and those directed to lie by an experimenter [20]. For example, those who chose to lie compared to those instructed to lie made fewer speech errors and hesitations, and fewer references to others. Therefore, studies utilizing real-world samples of subjects who either chose to be deceptive or

not may contribute more deeply to the understanding of deception than those studies utilizing mock lie scenarios, as well as provide more generalizable findings. Another criticism of mock lies is on the lack of motivation; participants have little to lose and do not chose to lie hence have little or no vested interested in whether or not they get caught [21]. A lack of personal involvement in the lie is another critique of laboratory studies [22].

Research has also shown the duration and content of a lie can influence how successful a person can be at deception. Longer lies, for instance, are more difficult to tell than short ones [23]. The idea that longer lies are more complex and difficult to maintain than short and simple lies is practically common sense. In the meta-analysis done by DePaulo et al., [12] they predicted that if deceivers were required to sustain their deception for greater lengths of time, then cues to deception would be clearer and more numerous. Their findings supported their hypothesis; duration did moderate the size of the effect. However, RWHS situations, like law enforcement interviews, may be longer than subjects will tolerate in a controlled experiment (e.g. the duration of the RWHS interview in this study lasted 14 hours over three days).

Although RWHS deception detection research could address these issues, it must overcome the wicked problem of establishing ground truth [24]. Ground truth is a verified or indisputable fact, for example adhering to evidentiary guidelines used in a court of law. In a laboratory setting, establishing ground truth is a matter of experimental design, fully controlled by the researcher. This same control is not possible in the real-world and to attempt to subject people to real stressors that would lead up to deceptive communication would be unethical and most likely illegal (e.g. ask a student to steal a computer from the schools lab and then monitor them during police interviews). In addition, random assignment of participants to treatment groups is not possible in field studies. Because of the wicked problem of establishing ground truth in RWHS deception and the unethical feasibility of laboratory experiments, case studies based on field data seem to be the experimental design with the greatest chance to further the understanding of deception detection.

Another issue with the body of deception detection research is that current theories lack adequate validation in RWHS settings [6], [25]. One promising theory that can benefit from validation in a RWHS is Interpersonal Deception Theory (IDT) [26] (Figure 1). According to IDT, the counterpart to senders' deception is receivers' suspicion. IDT suggests that deception is a dyadic interaction and as the deception takes place, receivers may become suspicious of the senders attempts to deceive and may adapt their behavior because of it.

For example, they may choose to conceal their suspicion by quickly moving on to another topic or admit their suspicion and confront the sender to gauge their reaction. While IDT already has empirical support [27], [28], examining it under a RWHS lens can strengthen its validation.

This study is an exploration of real-world, high-stakes (RWHS) deceptive behavior manifested in human speech, and analyzed by objective measures.

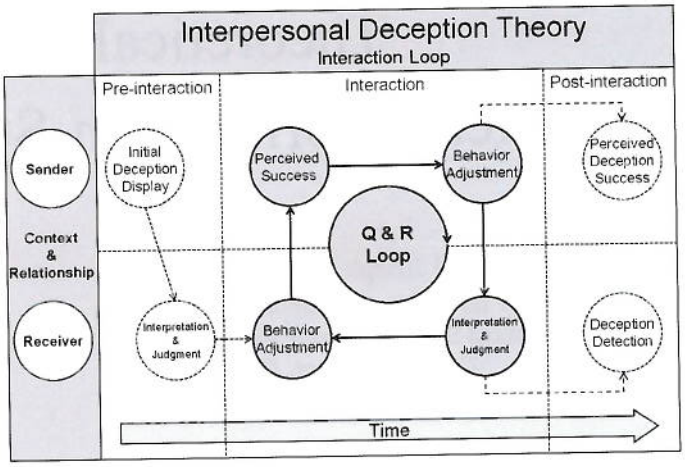

Figure 1, Interactive Deception Model (Adapted)

It is not a laboratory experiment with controlled settings in a closed environment, rather it is more akin to a field study. Though several statistical tools were employed and every opportunity to follow sound methodology was practiced, their use was not to prove/disprove hypothesizes but to explore the data and examine propositions based on theory, namely IDT. The impetus for this study came after a lengthy literature review on deception detection and has three tenets: (1) the state of existing theories on deception crave for validation, (2) outside the lab in a RWHS setting, (3) where typical dyadic interactions are long and more complex than those studied in a controlled setting. These tenets are the research gaps identified and where it is believed the most stands to be gained by exploration. In doing so, it serves to validate the phenomenon posited by IDT.

II. Methodology

Primarily, the study seeks to answer the question: Are speech cues to deceptive behavior moderated over time by receiver suspicion during dyadic interactions in a real-world high-stakes setting? In order to test this research question we first identified a communication channel that was easily measured with automated tools but also rich in behavioral cues. Speech is such a channel. Deception researchers have long been interested in speech as a source of behavioral cues [29], [30], [31], [32], [33]. Once the communication channel was identified, a list of cues and their constructs was chosen which previously research reported to be good indicators of deception.

Speech can be split into two categories, linguistic and paralinguistic. Linguistics is the study of what someone says and paralinguistics is how they say it. Our initial measures contained linguistic-based constructs from Fuller, Biros, and Wilson [34]. Fuller et al.'s study looked at 370 written suspect statements given during law enforcement interviews following criminal cases. Fuller et al.'s constructs and measurements were chosen because they generated almost 74% accuracy in deception detection, the data was RWHS field data taken in law enforcement environments with solid ground truth validation, and the units of measure were written statements.

This matches the current data set with the exceptions that it is a transcript of a law enforcement interview and the unit of measure varies from individual words to multiple sessions. The seven constructs used in this study are listed in Table 1.

Table 1, Linguistic Constructs & Measures

Construct	Construct Measurement	Brief Description
Quantity	# of Words, Verbs, & Sentences	Length of message
Specificity	Sensory ratio, Spatial ratio, Temporal ratio, Content Word Diversity, Bilogarithmic Type-Token-Ratio	Amount and type of details in the message
Uncertainty	Certainty Terms, Tentative Terms, Modal Verbs, Passive Voice, Generalizing Terms	Relevance, directness, and certainty of message
Clarity	Redundancy, Sentence Length, Complexity Ratio, Average Word Length, Causation Terms.	Message clarity and comprehensibility
Immediacy	1st person pronouns, 2nd person pronouns, 3rd person pronouns	Attempts to disassociate oneself from the events described
Affect	Activation, Imagery, Pleasantness*	Emotions present in the message
Cognitive Processing	Exclusive Verbs, Motion Words, Cognitive Processing Terms.	Increased or decreased cognitive processing and cognitive information present in the message related to veracity

* Note, Fuller [34] used positive and negative measures for each Affect measure, this study combines the positive and negative into a single bi-polar measure for ease of processing. In addition to the seven linguistic constructs by Fuller, listed above, an eighth construct of Severity was also considered by them to be important. However it is not a part of the current study because its measure would be constant across the current data set. The current data comes from a serial rapist, the punishment for which was life in prison. The lead detective in this case would assign the maximum severity score of five on the one to five scale used by Fuller.

For the paralinguistic measures we looked at the vocal constructs examined by Meservy [35]. These constructs and their measures were selected for this study because they represent a thorough coverage of the audio channel and tools exist to measure each. The six constructs were: Fluency, Duration, Tempo, Intensity, Frequency, and Voice Quality [29], [12]. However, because the construct Voice Quality contains cues that are difficult to measure objectively without the aid of a human evaluation this construct was removed; a focus of this study is on identifying behavioral cues that can be objectively measured and potentially automated. The five constructs and their 14 measures are described Table 2.

Table 2, Paralinguisit Constructs & Measures

Construct	Construct Measurement	Brief Description
Fluency	1. Non-ah disturbances 2. Speech errors 3. Silent pauses 4. Filled pauses	1. Speech disturbances other than "um", "er", "ah", and other such words 2. General speech errors 3, 4. Various pauses in conversation
Duration	1. Length of interaction 2. Response length 3. talking time	1. Total time of dyadic interaction 2. Length of sender's response 3. Proportion of total time sender talks
Tempo	1. Rate of speaking 2. Rate change	1. Average number of words per minute 2. Rate of speaking in the epoch minus the average rate of speaking for all responses
Intensity	1. Amplitude 2. Amplitude variety	1. loudness of senders voice 2. variation of loudness of a sender's voice
Frequency	1. Pitch 2. Pitch change 3. Pitch variety	1. The average fundamental frequency of sender's voice 2. variation of pitch of a sender's voice 3. Frequency of changes of pitch of a sender's voice

*note, the measures Interruptions from the construct Fluency and Response Latency from the construct Duration, are not considered due to the difficulty in automating these measures. Interruptions in the current study were removed because splitting speakers in a single channel audio recording is extremely difficult [36]. However, methods do exist for speaker-based segmentation which could be explored in future research [37].

Once a list of deceptive behavioral cues was identified we had to find a situation that met all the requirements of a RWHS dyadic interaction. There had to be a dyadic interaction between a sender and receiver whereby the sender might adjust his deceptive behavior when the receiver became suspicious of the sender's message. Fortunately, there is just such a RWHS that meets that criteria; the interview between an investigating police officer and a suspect in a criminal case. What follows is a description of the case.

III. Case Description

Please note, this case has been adjudicated and all identifiable information is publically available upon proper request. In Nov 2004, James Perry was sentenced in federal court in Madison, Wisconsin to 470 years in prison for creating child pornography, rape, sexual exploitation of children, child sexual assault and kidnapping; a crime spree that spanned over a five years and four states. It is the longest sentence for sex crimes in Wisconsin history and there is no option for parole.

In 2004 James Perry committed his final assault which led to his capture. Perry, a husband and father of two young girls, entered a Madison, Wisconsin hotel with the intent of committing a sexual assault against a 13 year old girl. This

incident was only one of two times Perry was ever caught on film despite targeting very public locations. It was a key piece linking him to a long series of rapes and assaults. At the same time the FBI was investigating a child pornography ring of which Perry was involved. Only a few days after the assault and attempted abduction of the young girl the FBI arrested Perry for his involvement in the internet child pornography ring.

The lead detective from the Madison Police Department (PD) became aware that the serial rapist she had been hunting was in FBI custody. The FBI was not aware of any rape or assault charges at that time. The lead detective informed the FBI about the plethora of crimes he committed and all plea bargaining on federal charges stopped so the lead detective could conduct the interview. The Madison PD had a list of 45 victims but believed there were hundreds more. Perry was highly motivated to lie because before the interview he was trying to proffer a plea agreement with the FBI for only a few years in prison in exchange for testifying against others in the child pornography ring. If additional charges for rape, sexual exploitation of children, child sexual assault and kidnapping were added, all plea bargaining would stop and he would be facing life in prison; an environment particularly not friendly to child molesters. Only after the interview and when Perry became aware of all the evidence against him was a plea agreement made to stop adding on charges (over and above the 125 he was now being charged with) because he was now most definitely going to prison for life.

Law enforcement videotaped three consecutive days of interviews totaling 14 hours and 27 minutes. Interviews were conducted by the same lead detective and her partner in the same room and under the same conditions with Mr. Perry and his attorney. Interaction was primarily between the lead detective and Mr. Perry, only minor contributions (less than five minutes total) were made by the second detective and Mr. Perry's attorney; their voices were removed before analysis. A 200 page law enforcement transcript was generated by the lead detective immediately after the interviews. The law enforcement transcript contains all questions asked and the responses, often in quotations with additional pertinent notes by the lead detective. Both the videotaped interviews and law enforcement transcripts were used in federal court. On the first day, the interview lasted just over four hours and 10 minutes, during which 711 individual questions were asked covering 209 different topics. The quality of the audio was very poor; a single microphone in a noisy room, typical for this setting.

Ground truth was established based on credible evidence admissible in a federal court. The lead detective identified four types of statements: (1) the truth, (2) suspected lies without evidence, (3) suspected lies with evidence, and (4) confirmed lies. Confirmed lies were those statements proven to be false by indisputable evidence admissible in court. When the sender made these statements law enforcement personnel knew for a fact he was lying. Suspected lies with evidence were those statements law enforcement personnel had disputing evidence on, however for various reasons that evidence was not or could not be admitted into federal court. Suspected lies without evidence were those statements law enforcement personnel believed, in their expert opinion, to be false but for which they had little or no evidence. The final type of statements are truthful, were the law enforcement personnel knew were the truth or had no reason to believe they were false.

Given a set of behavioral speech cues and their constructs, a clear definition of ground truth and a suitable RWHS data set, the next step was to run the analysis to determine if these cues were moderated over time by receiver's suspicion.

IV. ANALYSIS

The data preparation process followed the steps shown in Figure 2. First, the raw video stored on DVD was processed with Adobe Soundbooth to isolate the audio from the video portion; there was no loss of audio data during this step. The digital audio files were then passed through DC Live Forensic 7.5 to improve audibility in preparation for segmentation. Global filters were applied to remove audio signals outside the abilities of humans to hear as well as make. It should be noted that any filters or transformations to improve audibility were applied universally. It should also be noted that all recording took place in the same room with the same recording device and same environmental settings. Once global filters removed noise outside human speech range and audibility quality was improved, audio was segmented into question/response pairs and grouped by topic.

The audio was then duplicated for split processing for the two categories of cues, linguistic and paralinguistic. In preparing the audio for linguistic transcription any audio or acoustic filter can be applied that improves transcription accuracy (i.e. pitch, tone, cadence, etc. have no impact on linguistic cues). Linguistic cues were measured from the transcript using Structured Programming for Linguistic Cue Extraction (SPLICE) and Linguistic Inquiry and Word Count (LIWC) software. Waikato Environment for Knowledge Analysis (WEKA) (Witten & Frank, 2000) is used for classification based on the initial text processing steps. This transcript was then compared to the law enforcement transcript and deceptive statements were coded into the full transcript.

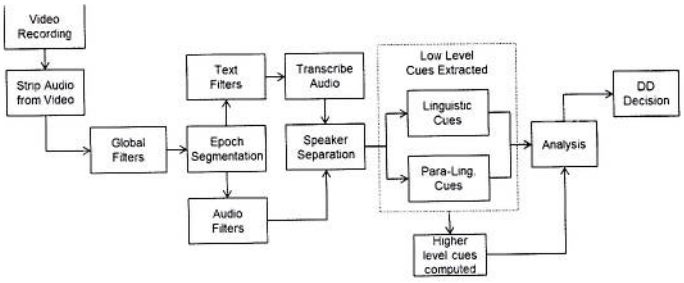

Figure 2, Data Processing Steps

The goal of processing the data for paralinguistic cue measurement is the removal of noise without removing, degrading, or changing the speech signal. There are several techniques for removing and improving clarity of audio however, some can be very aggressive and rely on human physical and cognitive audio processing characteristics to "trick" the listener into hearing clearer voices. This study took a conservative approach to audio filter selection to retain as

much of the voice signal as possible. DC Live Forensic 7.5 was the primary audio tool used for processing the paralinguistic measures.

The measures consist of 41 total measures across the 12 deception detection constructs. The linguistic-based cue constructs are: Quantity, Specificity, Uncertainty, Clarity, Immediacy, Affect, and Cognitive Processing. Paralinguistic-based cue constructs are: Time, Intensity, Frequency, Fluency, and Duration (Table 3). For this paper we pay particular attention to the linguistic construct Quantity and its' three measures, the number of words, verbs, and sentences.

Looking at Construct means required converting the individual measures to z-scores and averaging for each response. The following raw score mean tables give a good initial understanding of the spread of the data. For example, # of Words averaged just over 43 with truthful statements, less at 39 and deceitful statements, and much more at 66 words on average.

Table 3, Descriptive Means of Constructs

Constructs	Mean = 0 (Z-Score)			
	Truthful	↕	Deceitful	↕
Quantity	-0.096	↓	0.483	↑
Quantity	0.068	↑	-0.312	↓
Specificity	-0.018	↓	0.096	↑
Uncertainty	-0.034	↓	0.181	↑
Clarity	-0.028	↓	0.151	↑
Immediacy	-0.021	↓	0.138	↑
Affect	-0.049	↓	0.266	↑
Cognitive Proc.	-0.002	↓	0.013	↑
Fluency	-0.077	↓	0.398	↑
Duration	0.022	↑	-0.081	↓
Tempo	-0.013	↓	0.048	↑
Intensity	0.013	↑	-0.102	↓

Overall the mean scores for all but three constructs (75%) increased during deceitful behavior while truthful behavior showed a decrease in construct z-scores. Because all of the constructs are reflective (vs formative) it follows that changes in the individual cues reflect the changes in the latent constructs as seen in the following table [38], [39].

Table 4, Descriptive Means of Measures

Cues	Mean (Raw)	Truth	↕	Lie	↕
# of Words	43.72	39.25	↓	66.42	↑
# of Verbs	3.33	3.00	↓	5.03	↑
# of Sentences	8.31	7.41	↓	12.82	↑
Sensory ratio	0.79	0.81	↑	0.65	↓
Temporal ratio	4.77	4.71	↓	5.15	↑
Content Diversity	0.80	0.81	↑	0.73	↓
BTT-Ratio	79.37	80.74	↑	73.08	↓
Certainty Terms	3.022	3.03	↑	2.91	↓
Tentative Terms	3.111	3.07	↓	3.45	↑
Modal Verbs	10.49	10.24	↓	11.82	↑
Passive Voice	0.01	0.00	↑	0.00	↓
Gen. Terms	2.33	2.27	↓	2.44	↑
Redundancy	18.92	18.66	↓	20.32	↑
Sentence Length	12.66	12.53	↓	13.49	↑
Complexity Ratio	2.51	2.51	↓	2.51	↑
Avg Word len.	3.82	3.82	↑	3.78	↓
Causation Terms	1.03	0.85	↓	1.99	↑
1st p. pronouns	9.63	9.36	↓	11.29	↑
2nd p. pronouns	0.74	0.71	↓	0.77	↑
3rd p. pronouns	3.03	3.02	↓	3.11	↑
Activation	1.59	1.58	↓	1.66	↑
Imagery	1.40	1.39	↓	1.44	↑
Pleasantness	1.73	1.72	↓	1.78	↑
Exclusive Verbs	3.13	2.94	↓	4.09	↑
Motion Words	2.10	2.04	↓	2.50	↑
Cog. Proc. Terms	16.82	16.39	↓	19.25	↑
Non-ah distur.	2.30	2.19	↓	3.05	↑
Speech errors	0.010	0.010	↑	0.01	↓
Silent pauses	0.103	0.09	↓	0.11	↓
Filled pauses	2.010	2.05	↑	1.89	↓
Interaction len.	20.94	19.91	↓	26.33	↑
Response len.	13.00	11.78	↓	19.27	↑
Talking time	13.00	11.78	↓	19.27	↑
Rate of speaking	4.94	4.91	↓	5.06	↑
Rate change	0.65	0.69	↑	0.51	↓
Amplitude	53.72	53.68	↓	53.88	↑
Amp. variety	0.014	0.01	↓	0.014	↑
Pitch	135.1	136.5	↑	125.6	↓
Pitch change	0.053	0.05	↑	0.051	↓
Pitch variety	49.55	49.26	↓	51.05	↑

Overall 70.7% of the measures showed increases during deceptive responses reflecting a general rise in behavior measures (Table 4). This could be explained by deceiver's tendency to over compensate because he is anxious to appear

honest [34]. To better understand the differences between the group means, ANOVA was run on all constructs and measures.

An initial step to reporting ANOVA results should be to define what is "extreme". In other words, what is the cutoff value of α level of significance given the nature of the study. Most linguistic and psycholinguistic as well as MIS journals enforce the conventional α of 0.05 [40]. Because of the exploratory nature of this study Type II errors (failing to reject when the null hypothesis is in fact false) are more acceptable than Type I (rejecting the null hypothesis when in fact it is true). In practical terms, believing a treatment has an effect when in fact there is none (Type II error) is less damaging than dismissing a treatment that in fact has an effect (Type I error) [40]. Furthermore, given the uncontrolled environment from which the data was collected, a more relaxed α of 0.10 is adopted. The ANOVA of the 41 behavioral cues measured 29.3% as significant at the Q/R pair epoch level. Given the poor quality of the audio data, this is strong support for utilizing these measures in future deception detection research. What follows in Tables 5 and 6 are the ANOVA statistics on the constructs and individual measures z-score data.

In order to run the ANOVA on the constructs the data required manipulation so the aggregate of the different measures could be computed. All measures were given a z-score, on a positive scale across individual measures allowing for a meaningful average for each construct for each level of granularity.

Table 5, Construct ANOVA

CONSTRUCTS	By Topic	
	$F_{3, 707}$	Sig.
Quantity	1.527	.037
Specificity	1.440	.062
Uncertainty	0.621	.945
Clarity	1.208	.207
Immediacy	1.858	.004
Affect	1.222	.194
Cognitive Processing	1.478	.049
Fluency	1.494	.045
Time Duration	1.131	.289
Time Tempo	1.690	.013
Intensity	2.372	.000
Frequency	1.474	.050

ANOVA results on the constructs was very strong. There was a significant effect of Suspicion on 8 of 12 constructs ($0.621 < F3, 707 < 2.372$, $p < .062$). The strong performance of the paralinguistic constructs is encouraging if we consider the goal of automating the capture and processing of speech for deceptive measurement. All linguistic measures above syllable counting, require a speech recognition engine [41]. Given the poor quality of most RWHS audio recordings, paralinguistic measures may be the measures of choice in those environments.

Table 6, ANOVA by Granularity

MEASURES	Topic		MEASURES	Topic	
	F	Sig		F	Sig
NumSentences	2.14	.000	SilentPauses	2.37	.000
ContentWordDiv.	1.38	.087	AmpMeandB	2.40	.000
ComplexityRatio	1.72	.010	AmpVarietyPascals	2.27	.000
AvgWordLength	1.97	.002	PitchChange	1.64	.018
CausationTerms	1.68	.013			
1stppronoun	1.37	.090			
2ndppronoun	1.51	.040			
3rdppronoun	2.41	.000			
MotionWords	1.67	.014			

As seen in Table 6, there was a significant effect of Suspicion on the 13 of 41 measures at the Topic level of granularity ($1.37 < F3, 707 < 2.41$, $p < .09$).

To understand how the data behaved over time graphical analysis was performed and revealed promising results. The average magnitude for cues in each construct was graphed on a bar chart for side-by-side comparison. One example is given here, the Quantity construct clearly increase as Suspicion increases from Truth to Deception (Figure 3). However, the differences within the degrees of evidence are not clearly increasing. This may not be of concern with the exception of the w/o Evidence deception scores. One explanation for this maybe that the w/o Evidence level of suspicion had a very small sample size.

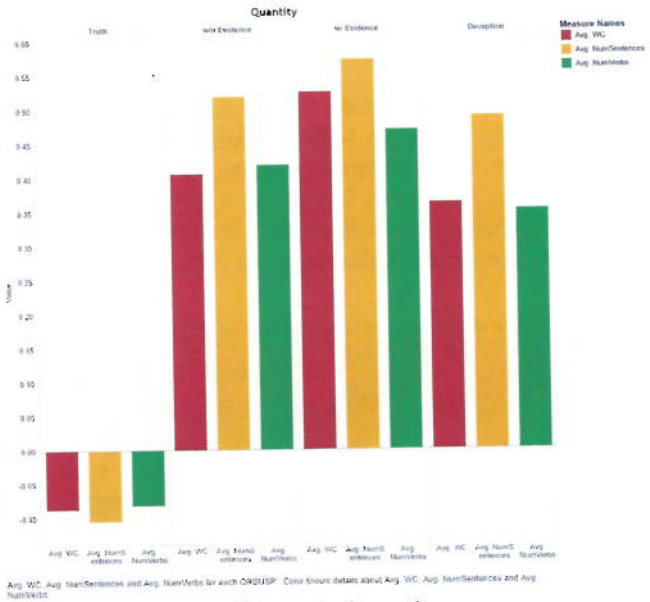

Figure 3, Quantity

In addition to bar graphs to examine magnitude, trend lines were drawn to examine general behavior over time. Figure 4 shows one example, again of the construct Quantity and its measures. It shows how Quantity decreases almost uniformly regardless of level of suspicion. One explanation for this pattern could be fatigue [42]. After four hours the subject could just be tired of talking. However, there is a stark difference in the Quantity spoken when comparing truthful vs deceptive speech which stay relatively constant, a pattern in and of itself.

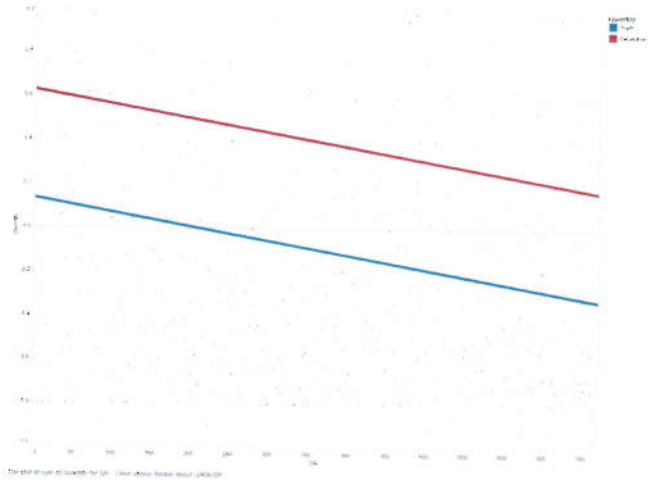

Figure 4, Quantity Line Graph

Figure 5 shows a comparison of truths (blue) to lies (red) over time for the three measures of the Quantity construct. Because these measures are highly correlated they behave similarly over time.

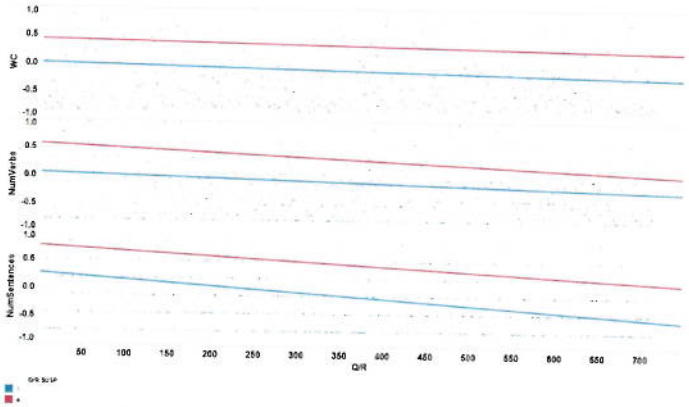

Figure 5, Quantity Trendlines

V. Conclusion

Based on the above it is reasonable to state that the research question was supported and that speech cues to deceptive behavior are impacted by receiver's suspicion during dyadic interactions in real-world high-stakes settings. This validated IDT in a RWHS setting. This study also look at whether measurements and constructs, developed by previous researchers, could hold up under a RWHS case. The ANOVA of the 41 behavioral cues measured 31.7% at the topic level. Regression also showed a strong relationship between the levels of suspicion and the individual measures with 31.7% at the Topic level of granularity as significant. Given the poor quality of the audio data, this is strong support for utilizing these measures in future deception detection research.

Several measures and constructs, utilized and validated in existing research, were explored and validated in this study. However, many of the measures and their constructs were not significant predictors of deceptive behavior or explained only a fraction of the variance. The reason for their poor predictive power could be explained because the study was a single case and the fact that all measurements were taken from an uncontrolled environment. However, this fact does add weight to those measures and constructs that were significant predictors of deceptive behavior.

In regards to IDT, one contribution of this study is a better understanding of the impact suspicion has in a RWHS setting. The length of the interaction in this case study was also a good opportunity to examine IDT and how a lengthy dyadic communication can be dissected into reasonable units of analysis. IDT was validated to the extent that suspicion plays a role in sender's behavior and it affected cue intensity. It is apparent that not only does suspicion play a central role in IDT but that its impact on deceptive speech behaviors is measurable in a RWHS environment. This point is important to unlocking future studies involving IDT, suspicion, and RWHS cases.

Several limitations are common to any case study. In the current study an emphasis was made to limiting research only to a RWHS environment, this raises a number of questions. Was this a typical high-stakes interview? Mr. Perry was more than a suspect, he knew the FBI had evidence against him, but he did not know how much evidence the lead detective had against him. Before the interview he wanted to plea down to six years for trafficking in child pornography; after the interview he received life in prison, 470 years to be exact. One could argue that having been caught, even on one criminal charge, he did not think he had much to lose by his deception.

The nature and environment of this real-world case is another limitation and potential area for further study. Longer, dyadic communication indicative of law enforcement interviews combined with a lack of fine granularity of episodes suggests the need for further research in interview-style communications. The difficulty is two-fold; longer duration interviews will be more difficult to gather in a controlled manner simply because volunteers are not going to sit for hours without proper compensation. Secondly, the free-flowing nature of longer communications makes controlling the study more complex.

The exploratory nature of the study, the volume of data, and the numerous methods of analysis used generated many possibilities for future research. One aspect of IDT which should be examined in greater detail is the view that deception involves strategic and non-strategic behaviors. This study's

initial view into a RWHS deceptive case did not look for strategic motives. However, such an examination could produce new insightful knowledge about deception, specifically in the case of longer more realistic dyadic interactions. This study kept IDT at the forefront when choosing the research question. However, there are several theories on deceptive behavior, all of which could benefit if looked at through a RWHS case study. One final potential future research area is the development of a collection of RWHS deception case studies. If a database of RWHS cases in which ground truth is established could be collected, it would be invaluable to the field of deception research.

VI. REFERENCES

[1] J. Hancock, Woodworth,and Goorha, "See No Evil: The Effect of Communication Medium and Motivation on Deception Detection," Group Decision and Negotiation, vol. 19, no. 4, 2010, pp. 327-343; DOI 10.1007/s10726-009-9169-7.

[2] J.T. Hancock, "Digital deception," *Oxford handbook of internet psychology*, 2007, pp. 289-301.

[3] C.F. Bond and B.M. DePaulo, "Accuracy of Deception Judgments," Pers*onality & Social Psychology Review (Lawrence Erlbaum Associates)*, vol. 10, no. 3, 2006, pp. 214-234.

[4] R.E. Kraut and D.B. Poe, "Behavioral roots of person perception: The deception judgments of customs inspectors and laymen," *Journal of Personality and Social Psychology*, vol. 39, no. 5, 1980, pp. 784-798; DOI 10.1037/0022-3514.39.5.784.

[5] A. Vrij, Edward, K., Roberts, K. P., and Bull, R., "Detecting deceit via analysis of verbal and nonverbal behavior," J. *Nonverbal Behavior.*, vol. 24, no. 4, 2000, pp. 239-263.

[6] A. Vrij, *Detecting lies and deceit : the psychology of lying and the implications for professional practice*, John Wiley, 2000.

[7] 1. G. Ganis, P. Rosenfeld, J. Meixner, R. Kievit, and H. Schendan,, "Lying in the scanner: Covert countermeasures disrupt deception detection by functional magnetic resonance imaging," *Neuroimage*, vol. 55, no. 1, 2011, pp. 312-319; DOI 10.1016/j.neuroimage.2010.11.025.

[8] D.D. Langleben, "Detection of deception with fMRI: Are we there yet?," *Legal and Criminological Psychology*, vol. 13, no. 1, 2008, pp. 1-9; DOI 10.1348/135532507x251641.

[9] G. Ganis, S.M. Kosslyn, S. Stose, W.L. Thompson, D.A. Yurgelun-Todd,, "Neural Correlates of Different Types of Deception: An fMRI Investigation," *Cereb. Cortex*, vol. 13, no. 8, 2003, pp. 830-836; DOI 10.1093/cercor/13.8.830.

[10] 1. R. Johnson, J. Barnhardt, J. Zhu, "The contribution of executive processes to deceptive responding," *Neuropsychologia*, vol. 42, no. 7, 2004, pp. 878-901.

[11] D.P. Twitchell, M. L. Jensen, J. K. Burgoon, and J. Nunamaker, "Detecting deception in secondary screening interviews using linguistic analysis," Proc. *The 7th International IEEE Conference on Intelligent Transportation Systems*, 2004. , 2004, pp. 118-123.

[12] B.M. DePaulo, et al., "Cues to deception," *Psychological Bulletin*, vol. 129, no. 1, 2003, pp. 74-118.

[13] A. Vrij and S. Mann, "Telling and detecting lies in a high-stake situation: The case of a convicted murderer," *Applied Cognitive Psychology*, vol. 15, no. 2, 2001, pp. 187-203.

[14] G. Ben-Shakhar and E. Elaad, "The validity of psychophysiological detection of information with the Guilty Knowledge Test: A meta-analytic review," *Journal of Applied Psychology*, vol. 88, no. 1, 2003, pp. 131.

[15] M.G. Frank and P. Ekman, "The ability to detect deceit generalizes across different types of high-stake lies," *Journal of Personality and Social Psychology*, vol. 72, no. 6, 1997, pp. 1429-1439.

[16] C.M. Fuller, K. Marett, and D. P. Twitchell, "An examination of deception in virtual teams: Effects of deception on task performance, mutuality, and trust," Professional *Communication, IEEE Transactions* on, vol. 55, no. 1, 2012, pp. 20-35.

[17] S. Porter and L. Brinke, "The truth about lies: What works in detecting high stakes deception?," *Legal and Criminological Psychology*, vol. 15, no. 1, 2010, pp. 57-75.

[18] L. ten Brinke and S. Porter, "Cry me a river: Identifying the behavioral consequences of extremely high-stakes interpersonal deception," *Law and Human Behavior*, vol. 36, no. 6, 2012, pp. 469.

[19] M.G. Frank and T.H. Feeley, "To catch a liar: Challenges for research in lie detection training," *Journal of Applied Communication Research*, vol. 31, no. 1, 2003, pp. 58-75.

[20] T.H. Feeley and M.A. deTurck, "The Behavioral Correlates of Sanctioned and Unsanctioned Deceptive Communication," *J. Nonverbal Behav.*, vol. 22, no. 3, 1998, pp. 189-204.

[21] G. Miller and J. Stiff, "Deceptive Communication. 1993," *Book Deceptive Communication*. 1993, Series Deceptive Communication. 1993, ed., Editor ed.^eds., Sage Publications, Thousand Oaks, CA, pp.

[22] R.J. Koper and J.M. Sahlman, "The Behavioral Correlates of Real-World Deceptive Communication," *Book The Behavioral Correlates of Real-World Deceptive Communication*, Series The Behavioral Correlates of Real-World Deceptive Communication, ed., Editor ed.^eds., 1991, pp.

[23] R.E. Kraut, "Verbal and nonverbal cues in the perception of lying," Journal of personality and social psychology, vol. 36, no. 4, 1978, pp. 380.

[24] W.G. Iacono, "Can we determine the accuracy of polygraph tests," *Advances in psychophysiology*, vol. 4, 1991, pp. 201-207.

[25] M. Hartwig and C.F. Bond, "Lie Detection from Multiple Cues: A Meta-analysis," *Applied Cognitive Psychology*, 2014, pp. n/a-n/a; DOI 10.1002/acp.3052.

[26] D.B. Buller and J.K. Burgoon, "Interpersonal deception theory," *Commun. Theory*, vol. 6, no. 3, 1996, pp. 203-242.

[27] D.B. Buller, K. D. Strzyzewski, and F. G. Hunsaker, "Interpersonal deception: II. The inferiority of conversational participants as deception detectors," *Communication Monographs*, vol. 58, no. 1, 1991, pp. 25 - 40.

[28] J.K. Burgoon, D. B. Buller, L. Dillman, J. B. Walther, "Interpersonal deception," *Human Communication Research*, vol. 22, no. 2, 1995, pp. 163-196.

[29] L. Anolli and R. Ciceri, "The voice of deception: Vocal strategies of naive and able liars," *J. Nonverbal Behav.*, vol. 21, no. 4, 1997, pp. 259-284.

[30] P.J. Fay and W.C. Middleton, "The ability to judge truth-telling, or lying, from the voice as transmitted over a public address system," *Journal of General Psychology*, 1941.

[31] D. Howard and C. Kirchhübel, "Acoustic Correlates of Deceptive Speech–An Exploratory Study," *Engineering Psychology and Cognitive Ergonomics*, 2011, pp. 28-37.

[32] J.A. Podlesny and D.C. Raskin, "Physiological measures and the detection of deception," *Psychological Bulletin*, vol. 84, no. 4, 1977, pp. 782-799; DOI 10.1037/0033-2909.84.4.782.

[33] M. Zuckerman, B. M. DePaulo, R. Rosenthal, and B. Leonard, "Verbal and Nonverbal Communication of Deception," *Advances in Experimental Social PsychologyVolume* 14, Academic Press, 1981, pp. 1-59.

[34] C.M. Fuller, D. P. Biros, R. L. Wilson,, "Decision support for determining veracity via linguistic-based cues," Decision Support Systems, vol. 46, no. 3, 2009, pp. 695-703; DOI DOI: 10.1016/j.dss.2008.11.001.

[35] T.O. Meservy, "Augmenting human intellect: Automatic recognition of nonverbal behavior with application in deception detection," *Ph.D. dissertation, The University of Arizona*, 2007.

[36] Z. Liu and M. Saraclar, "Speaker segmentation and adaptation for speech recognition on multiple-speaker audio conference data," Proc. *Multimedia and Expo, 2007 IEEE International Conference on, IEEE*, 2007, pp. 192-195.

[37] P. Delacourt and C.J. Wellekens, "DISTBIC: A speaker-based segmentation for audio data indexing," *Speech communication*, vol. 32, no. 1, 2000, pp. 111-126.

[38] M.G. Boltz, R. L. Dyer, A. R. Miller, "Jo are you lying to me? Temporal cues for deception," *Journal of Language and Social Psychology*, vol. 29, no. 4, 2010, pp. 458-466.

[39] T. Coltman, T. M. Devinney, D. F. Midgley, and S. Venaik, "Formative versus reflective measurement models: Two applications of formative measurement," *Journal of Business Research*, vol. 61, no. 12, 2008, pp. 1250-1262.

[40] K.R. Murphy, B. Myors, and A. H. Wolach, *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests*, Routledge, 2009.

[41] B. Bigi and D. Hirst, "SPeech Phonetization Alignment and Syllabification (SPPAS): a tool for the automatic analysis of speech prosody," *Proc. Proceedings of Speech Prosody*, 2012, pp. 1-4.

[42] G.M. Ramdharry, A. Thornhill, G. Mein, M. M. Reilly, and J. F. Marsden, "Exploring the experience of fatigue in people with Charcot–Marie–Tooth disease," *Neuromuscular Disorders*, vol. 22, 2012, pp. S208-S213.